[주제 3]

Logit analysis

- 모형 분석을 위한 기본 개념(1)
  - 자료 형태에 따른 분석 방법

		종속변수	
		비계량 (범주형/명목형)	계량
독립변수	비계량 (범주형/명목형)	Log-linear model	ANOVA
	계량	Logit Analysis Discriminant Analysis	Regression

  - Logit analysis (gr=2) = Logistic regression analysis
  - Log-linear model : 독립변수와 종속변수의 구분없이 변수들간의 연관성을 분석하는 방법(변수간의 연간성 및 교호작용 분석)

1. Logit Analysis
(Logistic regression)

- Basic principle
  - much the same as the (ordinary) multiple regression
  - main difference : to predict a transformation instead of a value of the dependent variable
    (to predict the transformation of a proportion)
- Recall : a binary variable
  - the categories numerical values of 0 and 1
    (yes or no)
  - mean=proportion

- Logit transformation
  - p ; the proportion of individuals with the characteristic
  - the ratio(=odds) and log odds

$$odds = \frac{p}{1-p}, \quad \text{logit}(p) = \log_e\left(\frac{p}{1-p}\right) = \ln\left(\frac{p}{1-p}\right)$$

- odds ratio
  - to compare predictions for subjects with or without a particular characteristic
  - the log of the odds ratio

$$\text{logit}(p_1) - \text{logit}(p_2) = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = \log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)$$

- Estimation of p : $p = \frac{e^l}{1+e^l}$, where $l = \text{logit}(p)$

1) 이항반응변수(a binary response variable)

- k explanatory variables : $X_1, X_2, \cdots, X_k$
  - Continuous, categorical (nominal and ordinal) data
- dependent variable : $y = \{1,\ 0\}$
  - Mean of y (success) : $E(y) = 1 \times p(Y = 1) + 0 \times p(Y = 0) = p$
  - Probability of y|x : $P(Y = y \mid X = x) = p^y (1 - p)^{1-y}$
- Probability model of success probability, p

$$p = \frac{\exp(\alpha + \beta\ x)}{1 + \exp(\alpha + \beta\ x)} \xrightarrow{s-curve} \frac{p}{1 - p} = \exp(\alpha + \beta\ x) (\Rightarrow \text{odds})$$

$$\xrightarrow{\text{Logit model}} \ln\left(\frac{p}{1 - p}\right) = \alpha + \beta\ x$$

  - odds increase as $\exp(\beta)$ as the explanatory x's increase
  - multiple logistic model : k explanatory variables

- Estimation of β
  - Maximum likelihood estimator(최대우도추정량)
  - Approximately Normal distributed
  - Test statistic under $H_0 : \beta_i = 0$

$$z_w = \frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)} \sim app. \quad \text{standard} \quad \text{Normal}$$

$$\chi^2_w = \left(\frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)} \right)^2 \sim app. \quad \chi^2(1)$$

  - If p-value < significance level then rejection

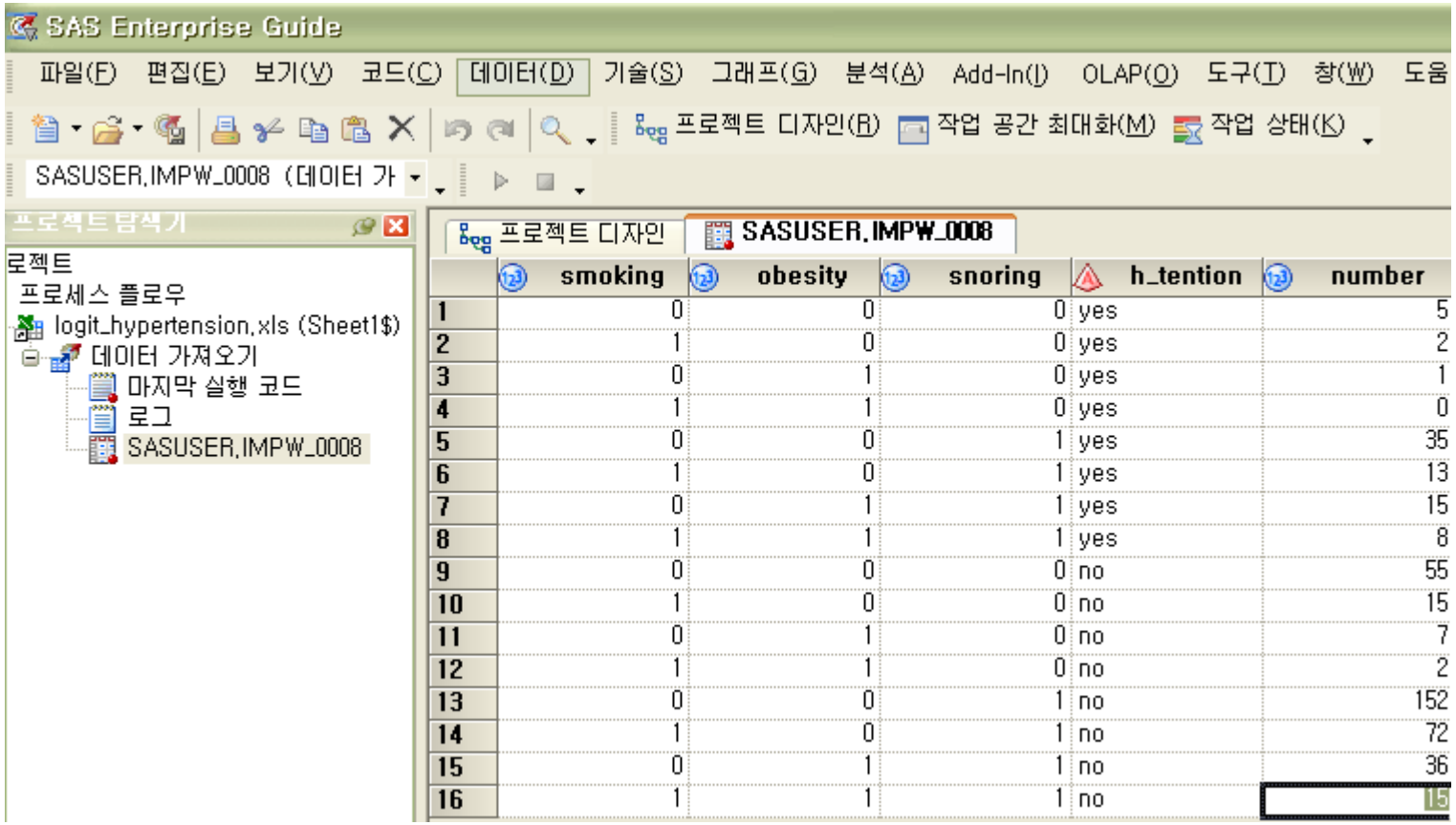
Example 1 : Hypertension

- Hypertension to smoking, obesity and snoring (433 men aged 40 or over, Norton and Dunn, 1985)

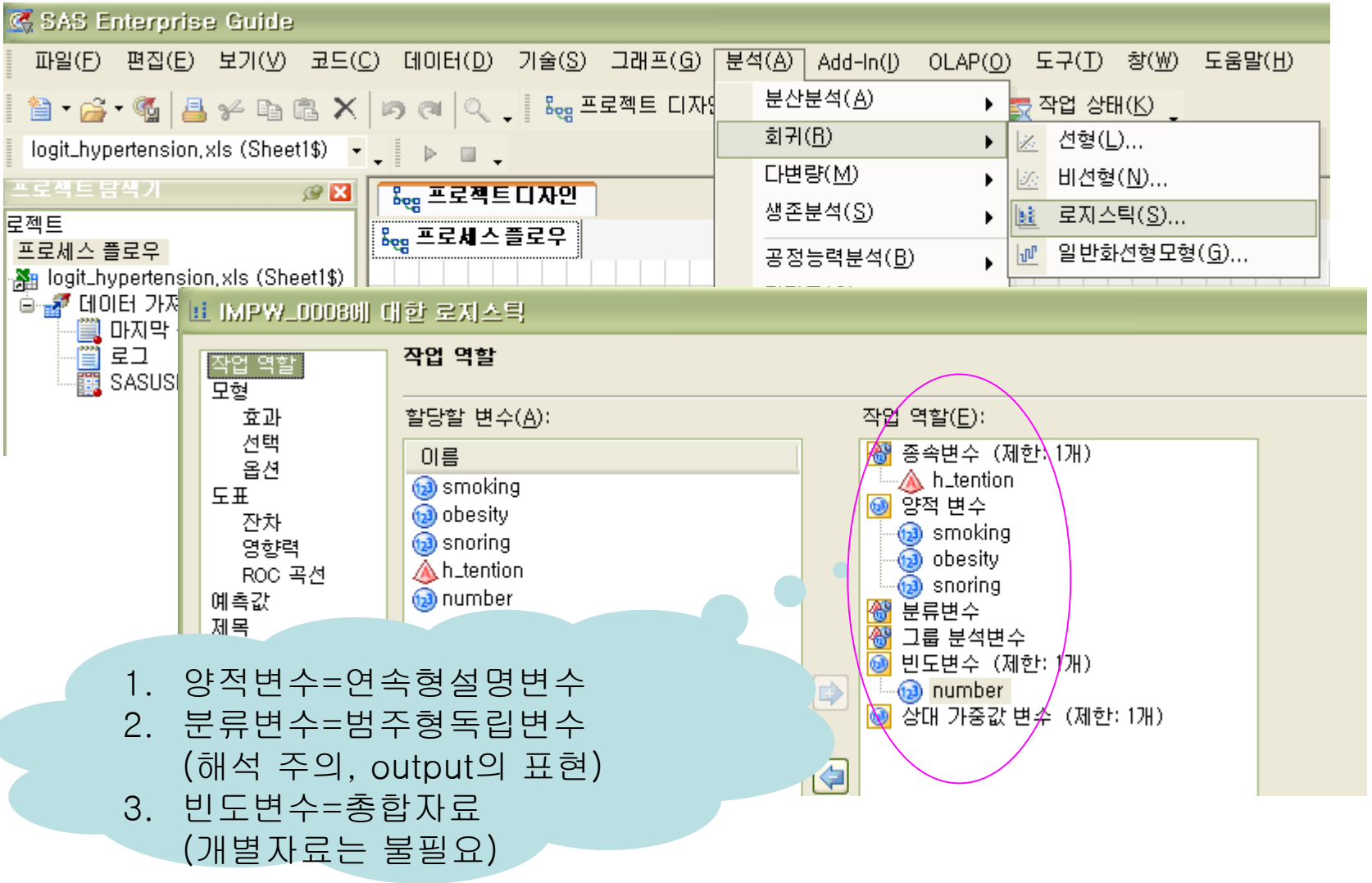
smoking	obesity	snoring	size	number	%
0	0	0	60	5	8.3
1	0	0	17	2	11.8
0	1	0	8	1	12.5
1	1	0	2	0	0.0
0	0	1	187	35	18.7
1	0	1	85	13	15.3
0	1	1	51	15	29.4
1	1	1	23	8	34.8
			433	79	18.2

  - to see which of the factors smoking, obesity and snoring are predictive of hypertension

- SAS에서 data 입력하는 방법



	smoking	obesity	snoring	h_tention	number
1	0	0	0	yes	5
2	1	0	0	yes	2
3	0	1	0	yes	1
4	1	1	0	yes	0
5	0	0	1	yes	35
6	1	0	1	yes	13
7	0	1	1	yes	15
8	1	1	1	yes	8
9	0	0	0	no	55
10	1	0	0	no	15
11	0	1	0	no	7
12	1	1	0	no	2
13	0	0	1	no	152
14	1	0	1	no	72
15	0	1	1	no	36
16	1	1	1	no	15



1. 양적변수=연속형설명변수
2. 분류변수=범주형독립변수
   (해석 주의, output의 표현)
3. 빈도변수=총합자료
   (개별자료는 불필요)

1. 주효과
2. 교차=교호작용

Model Information		
Data Set	WORK.SORTTEMPTABLESORTED	
Response Variable	h_tention	
Number of Response Levels	2	
Frequency Variable	number	number
Model	binary logit	
Optimization Technique	Fisher's scoring	

Response Profile		
Ordered Value	**h_tention**	**Total Frequency**
1	yes	79
2	no	354

Probability modeled is h_tention='yes'.

Yes를 성공확률로 사용

- 모형 적합성

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	413.424	408.567
SC	417.495	428.920
-2 Log L	411.424	398.567

- 유의성 검정
  (모든 회귀계수 추정치에 대한 검정)

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	12.8574	4	0.0120
Score	12.6809	4	0.0129
Wald	11.9542	4	0.0177

Wald 방법으로 분석 가능

- 개별회귀계수에 대한 추정치

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3504	0.3822	37.8274	<.0001
smoking	1	-0.1701	0.3309	0.2644	0.6071
obesity	1	0.5849	0.3429	2.9092	0.0881
snoring	1	0.8741	0.3976	4.8323	0.0279
smoking*obesity	1	0.3662	0.6176	0.3517	0.5531

➢ 로짓 모형식 :

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = -2.35 - 0.17\text{smoking} + 0.58\text{obesity} + 0.87\text{snoring} + 0.37(\text{smoking} \times \text{obesity})$$

- 유의한 변수에 대한 오즈비

Snoring의 오즈비

logit(p_s)-logit(p_ns)=

= e(0.8741) = 2.3967

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
snoring	2.397	1.099	5.225

1. 로짓모형식에서 유의한 snoring 의 오즈비를 보면, 코골이를 안하는 사람에 비해 코골이를 하는 사람들이 고혈압 환자일 오즈가 2.4배나 높다는 의미
2. 로짓모형식에서 유의하지 않은 주효과를 제외하고 최종 모형을 재적합 (교호작용 제거 후 주효과로만 재적합하여 주효과에 대해 판단)
3. 교호작용이 유의한 경우 주효과는 유의성에 관계없이 모형에 존재

- 연관성 측도

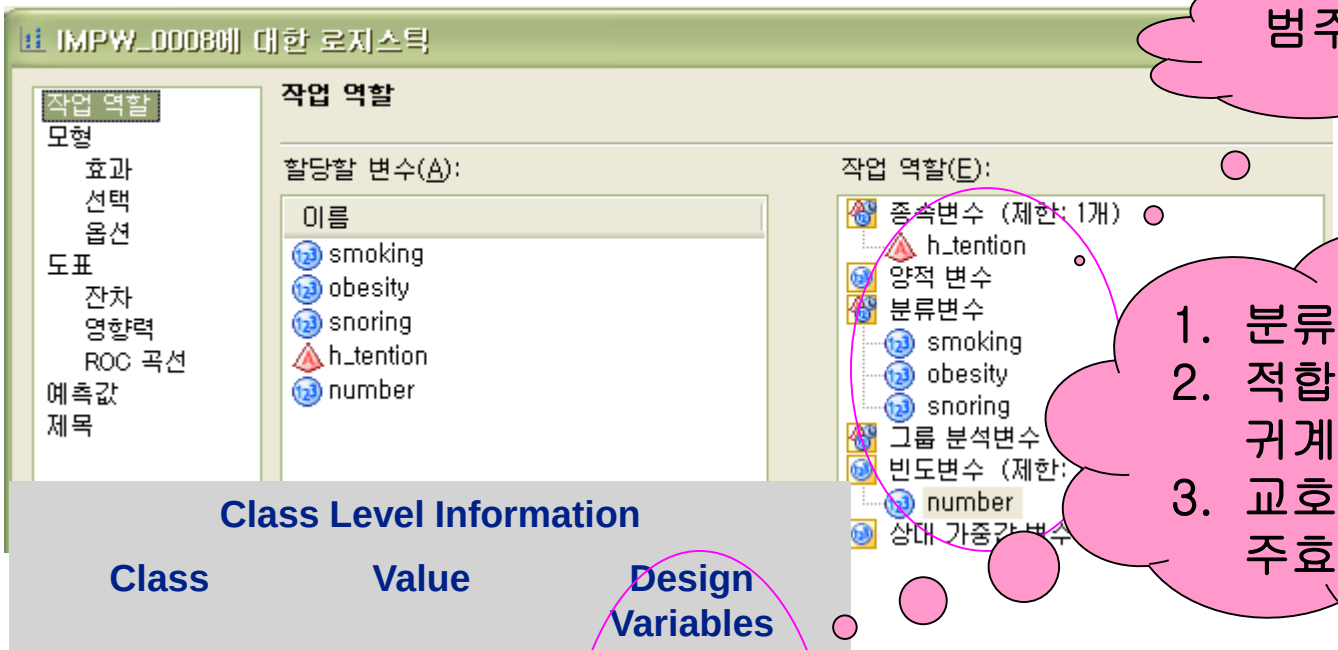
Association of Predicted Probabilities and Observed Responses

Percent Concordant	48.9	Somers' D	0.237
Percent Discordant	25.2	Gamma	0.320
Percent Tied	25.8	Tau-a	0.071
Pairs	27966	c	0.619

- 분류 결과

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.060	79	0	354	0	18.2	100.0	0.0	81.8	.
0.080	72	15	339	7	20.1	91.1	4.2	82.5	31.8
0.100	72	70	284	7	32.8	91.1	19.8	79.8	9.1
0.140	71	70	284	8	32.6	89.9	19.8	80.0	10.3
0.160	58	77	277	21	31.2	73.4	21.8	82.7	21.4
0.180	58	149	205	21	47.8	73.4	42.1	77.9	12.4
0.200	23	303	51	56	75.3	29.1	85.6	68.9	15.6
0.280	8	303	51	71	71.8	10.1	85.6	86.4	19.0
0.300	8	339	15	71	80.1	10.1	95.8	65.2	17.3
0.320	0	339	15	79	78.3	0.0	95.8	100.0	18.9
0.360	0	354	0	79	81.8	0.0	100.0	.	18.2



범주형 변수로 표현

1. 분류코딩에 주의
2. 적합성 및 모든 회귀계수검정 동일
3. 교호작용 제외한 주효과의 유의성

Class Level Information

Class	Value	Design Variables
smoking	0	1
	1	-1
obesity	0	1
	1	-1
snoring	0	1
	1	-1

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
smoking	1	0.0594	0.8075
obesity	1	5.9486	0.0147
snoring	1	4.8091	0.0283

- 개별회귀계수의 모수 추정

Type III과 동일

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6279	0.2227	53.4245	<.0001
smoking	0	1	0.0339	0.1391	0.0594	0.8075
obesity	0	1	-0.3477	0.1425	5.9486	0.0147
snoring	0	1	-0.4359	0.1988	4.8091	0.0283

➢ 로짓 모형식 :

$$\text{logit}\ (p) = \ln\left(\frac{p}{1-p}\right) = -1.63 + 0.03\text{smokin}\quad \text{g} - 0.35\text{obesit}\quad \text{y} - 0.44\text{snorin}\quad \text{g}$$

➢ **모형의 계수 해석에 주의**

1) 비만(-1)-정상(1)=0.35+0.35=0.70 : 비만이 정상보다 고혈압에 대한 로짓이 0.70배 만큼씩 증가한다고 해석
2) 비만 오즈 exp(0.35)-정상 오즈 exp(-0.35) : 비만여부의 오즈 exp(0.70)

- 모형의 주효과에 대한 오즈비

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
smoking 0 vs 1	1.070	0.620	1.846
obesity 0 vs 1	0.499	0.285	0.872
snoring 0 vs 1	0.418	0.192	0.912

➢ 오즈비 해석

1) 비만기준 오즈비 exp(-0.70)=0.497,
   정상기준 오즈비 exp(0.70)=2.014
2) 95% 신뢰구간에 "1"이 포함되면 주효과의 오즈비가 동일하다는 의미로 해석 : 비만여부 및 코골이는 유의미한 주효과로 판단

- 연관성 측도 및 분류결과는 거의 동일

2) 다항반응변수

- number of category of dependent variable : 3+
  - Extent to the binary logit model
  - Type : ordinal type and nominal type

A. Ordinal response variable

- generally use the proportional odds model
  - alternative model : adjacent-categories model, continuation-ratio model

B. Nominal response variable

- generally use the generalized logit model

A. Ordinal response variable : Proportional odds model

- response variable : J ordinal categories
  - prob. of jth response of ith subject : π_{ij}
  - cumulative prob. of jth response :
$$F_{ij} = P(Y \le j \mid \underline{x}_i) = \pi_{1j} + \pi_{2j} + \cdots + \pi_{ij}$$
- cumulative logit
$$L_j = \text{logit}\ (F_{ij}) = \ln\left(\frac{F_{ij}}{1-F_{ij}}\right) = \ln\left(\frac{\pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}}{\pi_{i\,j+1} + \pi_{i\,j+2} + \cdots + \pi_{iJ}}\right)$$
$$= \ln\left(\frac{P(Y \le j)}{P(Y > j)}\right)$$
- cumulative logit model
$$L_j(\underline{x}_i) = \alpha_j + \underline{\beta}'\underline{x}_i \quad \text{where} \quad \underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)'$$

- odds ratio of two value of explanatory variables with ith subject

$$\frac{P\left(Y \le j \mid \underline{x}_1^*\right) / P\left(Y > j \mid \underline{x}_1^*\right)}{P\left(Y \le j \mid \underline{x}_2^*\right) / P\left(Y > j \mid \underline{x}_2^*\right)}$$

- the proportional odds model

$$\ln\left(\frac{P\left(Y \le j \mid \underline{x}_1^*\right) / P\left(Y > j \mid \underline{x}_1^*\right)}{P\left(Y \le j \mid \underline{x}_2^*\right) / P\left(Y > j \mid \underline{x}_2^*\right)}\right)$$

$$= \ln\left(\frac{P\left(Y \le j \mid \underline{x}_1^*\right)}{P\left(Y > j \mid \underline{x}_1^*\right)}\right) - \ln\left(\frac{P\left(Y \le j \mid \underline{x}_2^*\right)}{P\left(Y > j \mid \underline{x}_2^*\right)}\right)$$

$$= \text{logit}\left(F_{1j}^*\right) - \text{logit}\left(F_{2j}^*\right) = \beta'\left(\underline{x}_1^*\right) - \beta'\left(\underline{x}_2^*\right) = \beta'\left(\underline{x}_1^* - \underline{x}_2^*\right)$$

  - the model is proportional to β

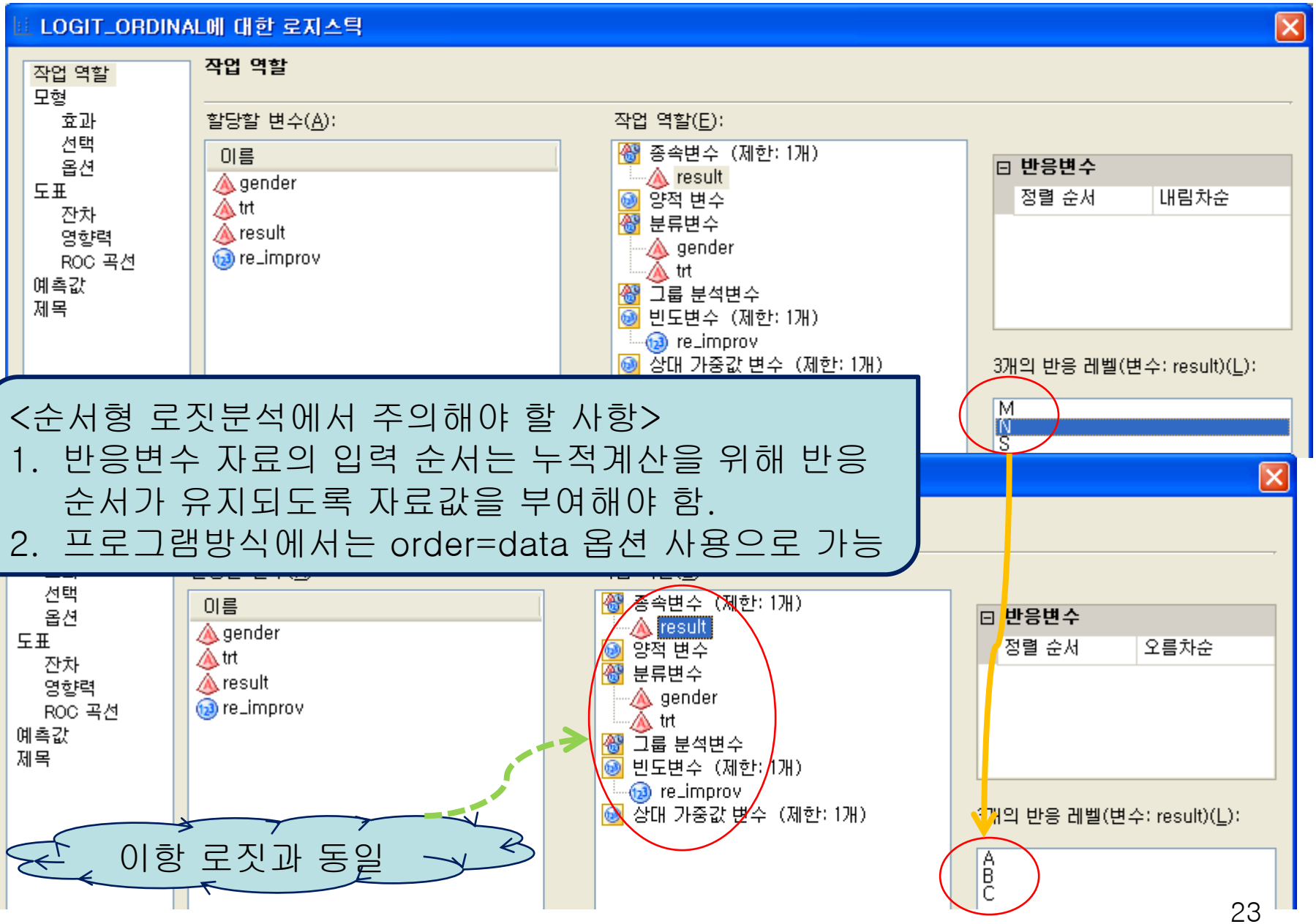
Example 2 : Arthritis pain

- Data : 3 ordinal responses of the improvement of arthritis pain with gender and treatments(3 types) (Kock and Edwards, 1988 : 교재 p.p. 79)

gender	trt	improve_N	improve_S	improve_M	total
F	A	6	5	16	27
F	P	19	7	6	32
M	A	7	2	5	14
M	P	10	0	1	11

사전분석 내용 : 여성, 처리 A의 개선 정도가 상대적으로 높게 나타남

- Goal : to analysis the effect of the improvement of arthritis pain


LOGIT_ORDINAL에 대한 로지스틱
작업 역할
모형
효과
선택
옵션
도표
잔차
영향력
ROC 곡선
예측값
제목
작업 역할
할당할 변수(A):
이름
gender
trt
result
re_improv
작업 역할(E):
종속변수 (제한: 1개)
result
양적 변수
분류변수
gender
trt
그룹 분석변수
빈도변수 (제한: 1개)
re_improv
상대 가중값 변수 (제한: 1개)
반응변수
정렬 순서
내림차순
3개의 반응 레벨(변수: result)(L):
M
N
S
<순서형 로짓분석에서 주의해야 할 사항>
1. 반응변수 자료의 입력 순서는 누적계산을 위해 반응 순서가 유지되도록 자료값을 부여해야 함.
2. 프로그램방식에서는 order=data 옵션 사용으로 가능
오름차순
A
B
C
이항 로짓과 동일

- Cumulative logit

Model Information		
Data Set	WORK.SORTTEMPTABLESORTED	
Response Variable	result	result
Number of Response Levels	3	
Frequency Variable	re_improv	re_improv
Model	cumulative logit	
Optimization Technique	Fisher's scoring	

Response Profile

Ordered Value	result	Total Frequency
1	A	42
2	B	14
3	C	28

Probabilities modeled are cumulated over the lower Ordered Values.

- 설명변수의 입력 정보

Class Level Information

Class	Value	Design Variables
gender	F	1
	M	-1
trt	A	1
	P	-1

- 비례오즈모형 가정의 검증

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
1.8833	2	0.3900

2(=3-1)개 반응범주의 비례오즈 로짓모형에서 설명변수의 효과가 동일한 가정에 대한 검정(채택)
- 기각 시 일반화된 로짓모형 사용

- 모형의 적합성 검증

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	2.7121	4	0.6780	0.6071
Pearson	1.9099	4	0.4775	0.7523

모형이 적합하다는 귀무가설을 채택
(참고 : 우도비 카이제곱=deviance)

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	173.916	158.029
SC	178.778	167.753
-2 Log L	169.916	150.029

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.8865	2	<.0001
Score	17.8677	2	0.0001
Wald	16.7745	2	0.0002

- 주효과에 대한 검정

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
gender	1	6.2096	0.0127
trt	1	14.4493	0.0001

- 효과의 모수 추정치에 대한 검정

반응 B이하와 C는 차이 있음을 보임

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	A	1	0.2547	0.2685	0.8999	0.3428
Intercept	B	1	1.1091	0.2949	14.1434	0.0002
gender	F	1	-0.6593	0.2646	6.2096	0.0127
trt	A	1	-0.8987	0.2364	14.4493	0.0001

➢ 로짓 모형식 :

$$A[\text{N Gr.}]: \text{logit}\ (F_{i1}) = 0.2547 - 0.6593 \times \text{Gender} - 0.8987 \times \text{trt}$$

$$B[\text{S Gr.}]: \text{logit}\ (F_{i2}) = 1.1091 - 0.6593 \times \text{Gender} - 0.8987 \times \text{trt}$$

- 오즈비의 해석 : 통증 개선 정도(No에 속할 가능성의 의미)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gender F vs M	0.267	0.095	0.755
trt A vs P	0.166	0.066	0.419

  - 성별 : 여성(1) -0.6593 + 남성(-1) 0.6593 = -1.3186

    통증 개선정도의 오즈비 exp(-1.3186)=0.267

    [해석] 통증 개선정도에 대한 성별의 오즈비가 0.267이라는 것은 여성이 No 그룹에 속할 가능성이 남성의 0.27배이므로 여성의 통증 개선정도가 남성에 비해 높다는 의미임.
  - 처리방법 : 처리의 오즈비가 0.166이므로 처리 A가 처리 P에 비해 No그룹에 속할 오즈가 0.17배로 낮다는 의미이며, 처리 A의 개선 정도가 높다는 의미임.

B. Nominal response variable : Generalized logit model

- response variable : J nominal categories
  - prob. of jth response : π_j where $\sum_1^J \pi_j = 1$

- generalized logit model : the base is Jth category

$$\ln\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \underline{\beta}_j' \underline{x}_i = \alpha_j + \sum_{l=1}^{k} \beta_{jl} x_{il} \quad (j = 1,2,\cdots, J-1)$$

$$\text{where} \quad \underline{\beta}_j' = (\alpha_j, \beta_{j1}, \beta_{j2}, \cdots, \beta_{jk})$$

$$\underline{x}_i = (1, x_{i1}, x_{i2}, \cdots, x_{ik})'$$

  - value of the generalized logit model = the base categorical logit

- Calculation of π_{ij} :

$$\pi_{ij} = \frac{\exp\left(\underline{\beta}_j' \underline{x}_i\right)}{1 + \sum_{k=1}^{J-1} \exp\left(\underline{\beta}_k' \underline{x}_i\right)} \quad \text{where} \quad j = 1, 2, \cdots, J-1$$

and

$$\pi_{iJ} = \frac{1}{1 + \sum_{k=1}^{J-1} \exp\left(\underline{\beta}_k' \underline{x}_i\right)}$$

Example 3 : Political party

- Data : the support of the political party
  (교재 p.p. 84)

Gender	Race	Party		
		Democ	Repub	Nopar
M	W	132	176	127
	B	42	6	12
F	W	172	129	130
	B	56	4	15

- Goal : to analysis the effect of the support of the party

- SAS 사용 시 주의
  - proc logistic : 이항형, 순서형 로짓분석 가능 (EG와 동일)
  - proc catmod/proc genmod : 일반화 로짓분석 가능 (EG에서 가능하지만 미해결)

```
data party;
  input gender $ race $ party $ count @@ ;
  datalines;
  M W Democ 132  M B Democ 42 F W Democ 172 F B Democ 56
  M W Repub 176  M B Repub 6 F W Repub 129 F B Repub 4
  M W Xpart 127 M B Xpart 12 F W X part 130 F B Xpart 15
;

proc catmod order=data;
  weight count;
  model party=gender race;
run;
```

- SAS output

The CATMOD Procedure

Data Summary

Response	party	Response Levels	3
Weight Variable	count	Populations	4
Data Set	PARTY	Total Frequency	1001
Frequency Missing	0	Observations	12

Population Profiles

Sample	gender	race	Sample Size
1	M	W	435
2	M	B	60
3	F	W	431
4	F	B	75

Response Profiles

Response	party
1	Democ
2	Repub
3	Xpart

1. 모형의 적합성 : 우도비 검정 결과를 이용하여 판단, 모형이 적합하다는 귀무가설 채택(p값 0.9056)
2. '주효과=0'에 대한 개별 효과 검정에서 효과가 없다는 귀무가설 기각

Maximum Likelihood Analysis

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	70.05	<.0001
gender	2	13.30	0.0013
race	2	57.89	<.0001
Likelihood Ratio	2	0.20	0.9056

Analysis of Maximum Likelihood Estimates

Parameter		Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	0.7190	0.1169	37.84	<.0001
		2	-0.4209	0.1902	4.90	0.0269
gender	M	1	-0.1101	0.0791	1.94	0.1641
	M	2	0.1763	0.0825	4.56	0.0327
race	W	1	-0.5591	0.1168	22.94	<.0001
	W	2	0.5799	0.1900	9.32	0.0023

➢ 로짓 모형식 :

$$\ln\left(\pi_1/\pi_3\right) = 0.7190 - 0.1101 \times Gender - 0.5591 \times Race$$

$$\ln\left(\pi_2/\pi_3\right) = -0.4209 + 0.1763 \times Gender + 0.5799 \times Race$$

- 범주 1(민주당)과 범주 2(공화당)의 차이에 대한 모형

  = 두 모형의 계수 차이를 이용하여 표현

$$\ln\left(\pi_1 / \pi_2\right) = 1.1399 - 0.2864 \times \mathrm{G}ender - 1.1390 \times Race$$

- 추정 오즈비의 계산

	민주당/무소속	공화당/무소속	민주당/공화당
성별 : 남성기준	Exp(-0.1101X2) = 0.8024	Exp(0.1763X2) = 1.4228	Exp(-0.2864X2) = 0.5639
인종 : 백인기준	Exp(-0.5591X2) = 0.3269	Exp(0.5799X2) = 3.1892	Exp(-1.1390X2) = 0.1025

  - 오즈비가 "1"보다 작음은 기준정당대비 비교정당의 지지도는 효과별 기준범주에 대한 비교범주의 지지도가 낮음을 의미함.

- 효과 수준별 반응 함수와 반응 확률의 계산
  - 예 : 남성(1)이고 백인(1)인 응답자의 정당별 추정 반응함수

$$[\text{M1 : Democ }] \ln\left(\pi_1 / \pi_3\right) = 0.7190 - 0.1101 \times \text{G}ender - 0.5591 \times Race = 0.0498$$

$$[\text{M2 : Repub }] \ln\left(\pi_2 / \pi_3\right) = -0.4209 + 0.1763 \times \text{G}ender + 0.5799 \times Race = 0.3353$$

  - 예 : 남성이고 백인인 응답자의 정당별 추정 반응 확률
    - 오즈비: 민주당 exp(0.1498)=1.0511, 공화당 exp(0.3353)=1.3984

$$\pi_{\text{남성 ,백인}} = \frac{\exp\left(\underline{\beta}_j' \underline{x}_i\right)}{1 + \sum_{k=1}^{J-1} \exp\left(\underline{\beta}_k' \underline{x}_i\right)} = \frac{\text{민주당 오즈}}{1 + \text{민주당 오즈} + \text{공화당 오즈}} = \frac{1.0511}{1 + 1.0511 + 1.3984} = 0.3047$$

$$\pi_{\text{남성 ,백인}} = \frac{\text{공화당 오즈}}{1 + \text{민주당 오즈} + \text{공화당 오즈}} = \frac{1.3984}{1 + 1.0511 + 1.3984} = 0.4054$$

$$\pi_{\text{남성 ,백인}} = \frac{1(\text{무소속})}{1 + \text{민주당 오즈} + \text{공화당 오즈}} = \frac{1}{1 + 1.0511 + 1.3984} = 0.2899$$

로짓 모형의 반응함수 값
- 예 1) 남성(1), 백인(1) , 민주당 : 모형 1 사용
- 예 2) 여성(-1), 흑인인(1) , 공화당 : 모형 2 사용

Maximum Likelihood Predicted Values for Response Functions

			------Observed------		------Predicted-----		
gender	race	Function Number	Function	Standard Error	Function	Standard Error	Residual
M	W	1	0.038615	0.124297	0.049772	0.119804	-0.01116
		2	0.326297	0.11643	0.335316	0.11483	-0.00902
M	B	1	1.252763	0.327327	1.16806	0.233468	0.084703
		2	0.69315	0.5	0.82453	0.378615	0.131382
F	W	1	0.27996	0.116216	0.269958	0.112389	0.010002
		2	-0.00772	0.124275	-0.01726	0.12224	0.009535
F	B	1	1.317301	0.29073	1.388246	0.229634	-0.07094
		2	-1.32176	0.562731	-1.1771	0.380534	-0.14465

Maximum Likelihood Predicted Values for Probabilities

			----Observed----		----Predicted---		
gender	race	party	Prob.	Standard Error	Prob.	Standard Error	Residual
M	W	Democ	0.3034	0.022	0.3047	0.0214	0.001
		Repub	0.4046	0.0235	0.4054	0.0233	-8E-4
		Xpart	0.292	0.0218	0.2899	0.0212	0.002
M	B	Democ	0.7	0.0592	0.6909	0.0442	0.0091
		Repub	0.1	0.0387	0.0942	0.0286	0.0058
		Xpart	0.2	0.0516	0.2149	0.0388	0.015
F	W	Democ	0.3991	0.0236	0.3978	0.0229	0.0013
		Repub	0.2993	0.0221	0.2985	0.0218	0.0008
		Xpart	0.3016	0.0221	0.3037	0.0216	-0.002
F	B	Democ	0.7467	0.0502	0.7539	0.038	-0.007
		Repub	0.0533	0.0259	0.058	0.0185	0.005
		Xpart	0.2	0.0462	0.1881	0.0347	0.0119

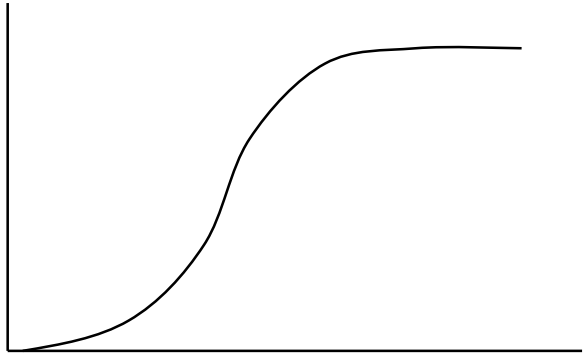
로짓 모형의 반응 확률
- 남성, 백인인 경우 민주당을 지지하는 반응 확률이 0.3047이라고 해석

[요약]
Logistic Regression/logit Analysis

- 독립변수가 다변량정규분포를 따르지 않은 경우에 유용한 방법
  - 독립변수가 이산형, 범주형, 연속형 등으로 이루어진 경우
- 일반적으로 종속변수가 이항형 자료
  - 종속변수의 분포 : 로지스틱함수로 가정(S형 곡선형태)

$$p = \Pr(Y = 1 \mid X = x) = \frac{\exp(\beta_0 + \sum_i \beta_i X_i)}{1 + \exp(\beta_0 + \sum_i \beta_i X_i)}$$

$$\Leftrightarrow \quad \text{logit}\ (p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_i \beta_i X_i$$



  - 범주의 수가 >2이상인 경우 : 일반화 로짓모형
  - 순서형 범주인 경우 : 순서로짓모형
- 로짓이 독립변수에 대해 선형함수라고 가정

- [Review 1] : odds의 의미

$$odds = \text{성공확률} \Big/ \text{실패확률} = p \Big/ 1-p$$

$$odds\ \ ratio = odd\ 1 \Big/ odd\ 2$$

$$= \Pr(Y=1 \mid X) \Big/ 1-\Pr(Y=1 \mid X)\ \text{, logistic Reg. and x is given}$$

- [Review 2] : 로지스틱회귀계수의 의미
  - $\exp(b)$ = X가 1단위 증가할 때

    Y = 1일 확률의 비가 증가하는 양 (odds ratio)
  - 오즈비는 사전확률에 영향을 받지 않으며, 자료의 모집단 분포가 달라도 추정 가능

- [Review 3] : 로지스틱회귀계수의 추정
  - 최대우도법으로 추정
  - 카이제곱 검정으로 변수의 유의성 검정

- [Review 4] : 입력변수가 질적 자료인 경우
  - 회귀분석과 같이 가변수로 사용
  - 가변수를 이용한 모형은 해석이 중요
    - 예 : 발병 여부(Y=1)에 대한 모형
    - 40대는 30대 미만 보다 exp(b1)배 만큼 사용
    - 50대 이상은 30대 미만보다 exp(b2)배 만큼 사용
    - 50대 이상은 40대보다 exp(b2)/exp(b1)배 만큼 사용

연령대	D1	D2
30대 미만	0	0
40대	1	0
50대이상	0	1

- [Review 5] : 변수의 선택
  - 회귀분석과 동일 : 전진선택, 후진제거, 단계별 선택
  - 오차제곱합 대신 로그우도함수값 사용
    - 변수 추가 시 로그우도값의 증가량이 가장 큰 변수 선택
  - 모형 선택 기준 : AIC 사용
    - AIC가 최대가 되는 모형으로 선택
- [Review 6] : 모형에 의한 분류방법
  - 주어진 독립변수에 대한 로지스틱 함수값 계산
  - 분류점(cut-off value) 설정 : 0< 분류점 <1
  - Pr(Y=1|X)>c 이면 집단 1로 분류

참고 : 분류분석(classification analysis)

1. 두 집단으로 분류되는 이유와 구분짓는 영향 변수는?
2. 적절한 판별(식별)함수로 집단 판별이 가능한 가?
3. 새로운 대상의 식별(집단 예측)이 가능할 까?
4. 집단 분류 오류 발생 가능성 존재

분류분석의 기본 개념

- 분류 분석이란?
  - 측정 자료로부터 집단을 식별하여 분류
  - 집단을 모르는 경우 : 집단 분류가 분석 목적
    (군집분석이 널리 이용)
  - 집단이 알려진 경우 : 집단 분류변수의 탐색, 분류함수의 추정
    새로운 대상의 소속 집단 예측 등이 목적
    (판별분석, 로지스틱회귀분석, 로짓분석 등)

- 집단 분류(식별 혹은 판별)의 응용 사례
  - 금융권에서 대출 여부를 판별하는 모형
  - 발병여부에 대한 판별
  - 쿠폰 혹은 카드 사용여부에 대한 판별
  - 도시특성에 대한 판별
  - 계속 이용여부에 대한 판별

판별분석과 로지스틱 회귀분석의 기본개념

- 공통점 : 집단분류분석방법으로 널리 이용
  조건부 확률의 추정에 기반한 모형

- 판별분석
  - 측정 자료를 이용하여 개별 개체들이 2개 이상의 집단 중 특정한 하나의 집단에 속하는 지 판별하는 다변량분석방법
  - 군집분석과의 차이점 : 집단 결과에 대한 사전 인지 여부

- 로지스틱 회귀분석(로짓분석)
  - 반응(종속)변수가 '0'과 '1'과 같은 이산형 혹은 가변수인 경우에 이용되는 모형
  - 독립변수가 다변량정규분포를 따르지 않는 경우 판별분석을 대체하는 유용한 분석 모형
  - 회귀분석과의 차이점 : 종속변수의 형태(이항형, 로지스틱 함수)